A DNA-based, Biomolecular Cryptography Design

Jie Chen
Division of Engineering, Brown University, RI 20912, USA

Abstract

Recent research has focused on DNA as a medium for ultra-scale computation and for ultra-compact information storage. Unlike using the conventional electronic means to encode data, we propose a novel design of DNA-based, molecular cryptography design in this paper. Due to the random nature of DNA, we can make our cryptography in principle unbreakable. Carbon nanotube-based message transformation, and DNA-based cryptosystem are proposed. To demonstrate the performance, we present an interesting example to encode and decode images using the proposed scheme.

I. Introduction

In a pioneering study, Adleman used DNA to solve a directed Hamiltonian path problem [1], thereby demonstrating the feasibility of a molecular approach to solve combinatorial problems. This approach has been extended by Lipton to solve another NP-complete problem [2] – the "satisfaction problem. These elegant studies demonstrated how problems corresponding to Boolean formulas can be solved by a massively parallel processing procedure. Such a procedure makes use of the ability of DNA sequences to hybridize specifically to their complementary sequence. Recently, much of attention has been drawn to next-generation data storage media. DNA and RNA are appealing media for data storage due to very large amounts of data that can be stored in compact volume. They far exceed the storage capacities of conventional electronic, magnetic and optical media. For instance, a gram of DNA contains about 108 tera-bytes. Hence, a few grams of DNA may have the potential of storing all the data stored in the world. In [3], the authors demonstrate the use of an AC electric field to address DNA used as a data storage media.

Once we can store data using DNA, data security and cryptography become the critical aspects to protect stored data. In cryptography, a fixed codebook may be used to provide an initial mapping from characters in the finite plaintext alphabet to a finite alphabet of codewords, and then a sophisticated algorithm depending on a key may be applied to further encrypt the message. Decryption (also known as decipherment) is the reverse process of transforming the encrypted message back to the original plaintext message. A cryptosystem (also known as a cipher) includes methods for both encryption and decryption of data. "Classic" encryptions, which are primarily text based, treat data as a character string or a bit stream [4] and generally do not make use of inherent structures or the syntax of the data. A cipher can be a block or a stream cipher. A block cipher applies encryption to a block of data at once, and requires significant computation. The Digital Encryption Standard (DES) and the RSA public-key encryption are two examples of block cipher [5]. A stream cipher typically performs bit-wise exclusive-or on each bit of the plain text with a one-time-pad [5]. Unlike conventional electronic approaches, we investigate a *biomolecular* method for encrypting and decrypting data in this paper.

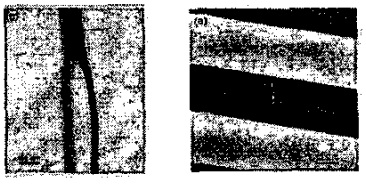
The processes of encryption and decryption requires:

(1) an informational message to be transmitted;
(2) a physical means for message transformation; and
(3) agreed encryption and decryption algorithms.

We organize the paper to address these three tightly related issues. With respect of issue one, we assume the plaintext message data is encoded in DNA strands using the publicly known alphabet of short oligonucleotide sequences. For instance, DNA strands in the solution of a test tube may form a *"wet"* data base that needs to be secured. In Section II, we briefly discuss our carbon nanotube-based probe to transform data between DNA and conventional binary storage media. Then, we present the proposed DNA cryptography mechanism in Section III. The paper is finally concluded in Section V.

II. Carbon Nanotube-based Probe

Our proposed nano-scale probe is centered around the use of Y-junction carbon nanotubes (CNT) for our DNA encoded data write-in/read-out. This nano-scale device allows us to access or alter information at the atomic scale. The Y-shaped CNTs are produced by CVD growth in branched nanochannel alumina templates [6]. This method produces multi-walled CNTs in aligned arrays with adjustable *stem* and *branch* tube diameters as shown in Fig. 1 (a). The Y-shaped CNT is typically 6 to 10 μm long as shown in Fig. 1 (b) and tubes with stem/branch diameter ratios of approximately 60nm/40nm are typical. The Y-junction CNT is



(a) (b)

Figure 1: Y-junction carbon nanotubes. (After Papadopoulos *et al.* [7]) (a) Transmission electron micrography of a Y-shaped CNT removed from the template. (b) Scanning electron micrograph of a 6μm long Y-shaped tube across gold electrodes.

a very promising element because it can act as both a

0-7803-7761-3/03/$17.00 ©2003 IEEE

connection and as an active device. Moreover, there exists a reasonably extensive toolkit of experimental device results on which to base our component designs[6],[8],[9]. Another attractive characteristic of the Y-CNT is that the Y junction can act as a pair of diodes. The diode properties of the Y-CNT arise because the diameter of nanotube decreases at the junction of the Y. Both the stem and the branch are p-type semiconductors, but the band-gap in the smaller diameter branch is larger than the band-gap of the stem as illustrated in Fig. 2(a). These diodes are very useful in constructing selection and switching circuits that can be incorporated into memory cell arrays. The branching structure of the

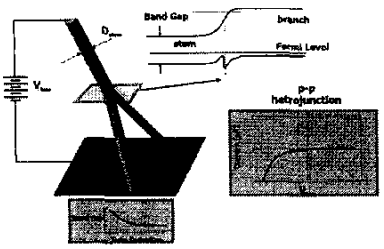


Figure 2: The band-gap of a CNT varies in a continuous manner with tube diameter, as shown on the lower left. The Y nanotube on the left forms a heterojunction at the branch point, as illustrated by the band structure in the upper right. The I-V characteristics of the diode is shown on the lower right (After Li [6]).

Y-CNT is an essential element of our proposed nano-scale probe. The branching provides a natural fan-in and fan-out mechanism. Connecting the stem of one Y-CNT to the branch of a preceding Y-CNT enables a significant fan-out with relatively few inter-tube connections. An arrangement using vertical stacking is illustrated in Fig. 3 to write-in/read-out encoded data in DNA strands.

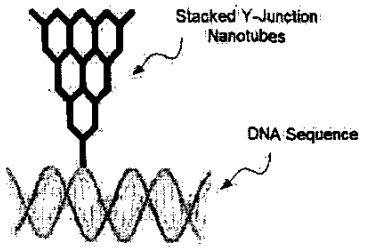


Figure 3: Nano-scale probe.

III. Biomolecular Cryptosystem

We assume inputs to our DNA cryptosystems are short segments of distinct plaintext messages. One-time-pad encryption uses a random *codebook* to convert short segments of plaintext messages to encrypted text. Construction of the libraries of codebook pads can be achieved using segmental assembly or build-up procedures used successfully in gene library construction projects [10]. Two critical points should be made about codebooks – they must be truly *random*, and they must be used *only once*. Such a cryptosystem using a secret random one-time-pad is the only cryptosystem known to be absolutely unbreakable [5]. We first assemble a large one-time-pad in the form of a DNA strand, which is randomly assembled from short oligonucleotide sequences. Then the DNA strand is isolated and cloned. These one-time-pads are assumed to be constructed secretly. Furthermore, we assume that specific one-time-pads are shared in advanced by both the sender and receiver of the secret message.

A. DNA Cryptography

Our DNA encoded messages are manipulated such that the plaintext strands are converted to cipher strands by adding message words to one-time-pads.

A.1. Construction of DNA Cipher Words

There are a number of possible methodologies for construction of cipher words used for the cryptosystems. One methodology is the random assembly of one-time-pads in solution. We view such methods less favorably due to the difficulty of achieving both full coverage and yet still avoiding possible conflicts by repetition of plaintext and/or cipher words. Our favorable method is to employ a DNA chip technology [11], [12]. Such DNA chips are currently commercially available; and chemical methods for construction of custom variants are well developed. The DNA chip has an array of immobilized DNA strands, so that multiple copies of a single sequence are grouped together in a microscopic array. These microscopic arrays of the DNA chip are carbon nanotube (CNT) probe addressable. There is known technology for growing distinct DNA sequences at each site of the array. DNA synthesis can be conducted in parallel. Therefore, the number of sequences prepared far exceeds the number of chemical reactions required. For preparation of oligonucleotides of length L, the $4L$ sequences are synthesized in $4n$ chemical reactions. For example, the 65,000 sequences of length 8 require 32 synthesis cycles. The plaintext and cipher pairs can be constructed so that there is a nearly unique word mapping between plaintext and cipher pairs. These resulting cipher word, plaintext word pairs can be assembled together in random order on a long DNA strand by a number of known methods, e.g., simply blunt end ligation. For higher efficiency, the use of a technique of hybridization assembly with complemented pairs has been done in Adleman's original DNA experiment [1]. Cloning or polymerase chain reaction (PCR) may be used to amplify the resulting one-time pad. Again, the cipher words should be constructed so that there is nearly complete coverage for each pad and a nearly unique cipher word.

A.2. DNA-based Addition

A general algorithm for DNA-based modulo-2 addition of any two non-negative rational binary numbers is presented. We begin with the DNA representation of all possible pairs of input non-negative binary bits as shown

in Figure 4. The "first bit" and "second bit" at a given

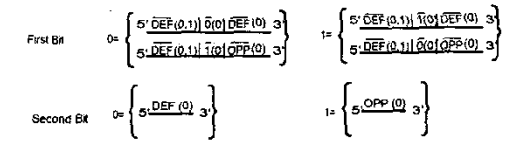


Figure 4: DNA representation of all possible non-negative binary bits to be added.

position refer to the value of either 0 or 1 of bit to be added in the original message and in the cipher word, respectively. DNA sequences are single-stranded, unique, and non-complementary, except that overlining indicates a complementary DNA sequence, for example, $\overline{DEF}(0,1)$ is complementary to $DEF(0,1)$. A number in parentheses refers to a position, whereas a number not in parentheses refers to the value of the digit at that position. Here the position information provides CNT-probe accessibility for data write-in/read-out. The first bit is represented by two DNA strands, each containing (from the 5' end) a "position transfer operator". The second digit is represented by a single DNA strand with the sequence DEF or OPP if the digit is either 0 or 1, respectively. This strand represents an operator that serves as a primer in a primer extension reaction.

As a schematic example, we illustrate how to add binary $1 + 1$. The bio-chemical reaction is shown in Figure 5. In the reaction, the operator (primer) represent-

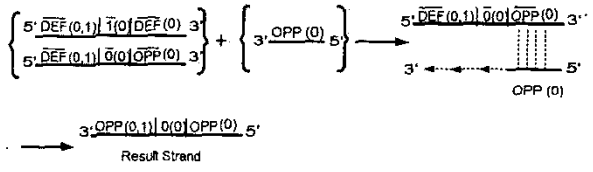


Figure 5: Illustration of the operation $1 + 1$ as an example of a DNA-based addition of two binary bits. Vertical dotted lines represent hybridization between complementary DNA elements, and reiterated arrows represent primer extension.

ing the bit i in original message sequence (W_i) hybridizes to the appropriate DNA strand representing the coded bit i (C_i). Primer extension yields result DNA strand. This successive reaction represents an example of a process of horizontal chain reaction, in which input DNA sequences serve as successive templates for extension of result strands.

A.3. Cryptography Mechanism

In classical cryptography [13], the Vernam cipher (now known as the XOR one-time-pad cryptosystem) is deployed by generating a sequence, S, of R independently distributed random bits known as a one-time-pad. The one-time-pad is replicated, and stored one copy at the source and one at the destination. Let L be the number of bits of S that remain unused, where initially $L = R$. Recall that XOR is the operation that essentially is Boolean modulo-2 addition discussed previously. When a plaintext binary message M which is $n < L$ bits long needs to be sent, each bit M_i is XOR'ed with the message bit K_i to produce encrypted bits

$$C_i = M_i + K_i \quad \text{for i =n.}$$

The n bits of S that have been consumed are then destroyed at the source. The encrypted sequence $C = (C_1, C_2, \ldots, C_n)$ is then despatched to the destination. At the destination, the identical process is repeated. The sequence C is used in the place of M. We perform bitwise modulo-2 addition with bits from S, and then the bits of S are destroyed after they are consumed. The commutative property of the addition results in the initial message being reproduced since $C_i + K_i = M_i$. In principle, we wish to convert one test tube of DNA strands (the plaintext messages) into another set of entirely different strands (the encrypted messages) in a random yet reversible way. Each of the plaintext messages are assumed to have appended unique prefix index tags of fixed length L_0. Each of the one-time-pad DNA sequences are also assumed to have appended unique prefix index tags of the same length L_0 that form the complements of the plaintext message tags. By use of known recombinant DNA techniques (e.g., annealing and ligation), each corresponding pair of a plaintext message and a one-time-pad sequence, with the same tag, can be combined into a single DNA strand. Our DNA encoded messages are modified in this case by a bit-wise modulo-2 addition computation, so that fragments of the plaintext are converted to cipher strands using the one-time-pad DNA sequences, and the plaintext strands are removed afterwards.

A.4. Parallel DNA Cryptosystem

In order to take advantage of the massive parallel processing capabilities of biomolecular computation, the following method for basic operations such as arithmetic (addition and subtraction) permit chaining of the output of these operations into the inputs to further operations, and to allow operations to be executed in massive parallel fashion.

Generalization of the addition to two non-negative vectors with n-digit binary numbers is straightforward. The two bits in each of the positions 2^0 through 2^n are represented as shown in Figure 4 with the following modification. At a position i other than 1, unique DNA sequences represent the values $0(i)$ and $1(i)$, and operators are replaced appropriately; for example, $\overline{DEF}(1)$ and $\overline{DEF}(1,2)$ by DNA sequences representing $\overline{DEF}(i)$ and $\overline{DEF}(i, i+1)$. The modulo-2 addition operation is in theory exactly as described above. This operation yields a final result strand longer than that shown in Figure 5 but with the same basic structure. This more general algorithm can readily be extended to the addition of any two n-digit positive rational numbers.

We have designed and executed an application of DNA-based addition of all possible pairs of non-negative binary vectors. Addition is performed by combining a test tube primer extension reagents plus DNA one-time-pad strands that appropriately represent two numbers

to be added. The addition is then followed by a primer extension reaction. The predicted reactions are illustrated in Figure 6. For $0+0$, the 20-base input operator

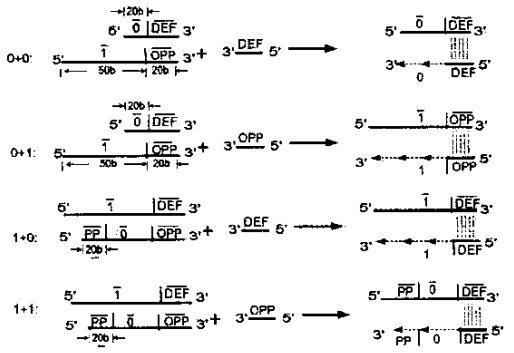


Figure 6: Illustration of a general DNA-based algorithm for adding two binary digits.

DEF hybridizes specifically to the 40-base input strand. Primer extension elongates the DEF operator to yield a 40-base result strand that encodes DEF and the result 0. For $0+1$, the 20-base input operator OPP hybridizes to the 70-base "opposite" strand. Primer extension yields a 70-base result strand that encodes OPP and the result 1. Similarly, input of $1+0$ yields a 70-base result strand encoding DEF plus the result 1. For $1+1$, the first primer extension reaction yields the 60-base result strand and transforms the potential primer creator $\overline{PP}$ into the actual primer PP. This is the simplest example of the horizontal chain reaction. For modulo-2 addition, since we don't care about the carrier, the addition takes $O(1)$ to combine two sequences with length of n in a massive parallel fashion.

IV. Simulation

The previously mentioned theoretical algorithm was executed biochemically as shown in Figure 7. The experiment shows that the $0+0$ operation yields the expected 40-base-long result strand, whereas the operations $0+1$ and $1+0$ each yields the expected 70-base-long result strands, thereby satisfying biochemically the required commutativity of addition. In Figure 8, we present an interesting example of using the DNA cryptography technique to encode a two-dimensional image.

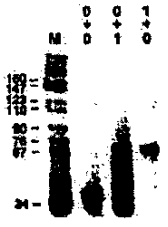
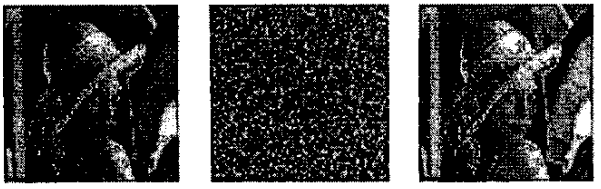


Figure 7: Biochemical execution of a simple DNA-based algorithm for adding two binary bits. M is a single-stranded molecular size marker. Molecular sizes are indicated on the left (in bases).



(a) Original (b) Encrypted (c) Decrypted

Figure 8: Simulation of the proposed DNA cryptography.

V. Conclusion

DNA and RNA are appealing media for data storage due to very large amounts of data that can be stored in compact volume. They far exceed the storage capacities of conventional electronic, magnetic and optical media. In this paper, we present a novel DNA-based cryptography technique that takes advantage of the massive parallel processing capabilities of biomolecular computation. We secretly assemble a library of one-time-pads in the form of DNA strands. Then, a modulo-2 addition method is employed for encryption whereby a large number of short message sequences can be encrypted using one-time-pads. We also present an interesting example of encrypting/decrypting a two-dimensional image.

References

[1] L. M. Adleman, ", *Science*, , no. 266, pp. 1021–, 1994.

[2] R. J. Lipton, ", , no. 268, pp. 542–, 1995.

[3] W. A. Germinshuizen, "Data storage using DNA", in *10th Foresight Nanotechnology Conference*, 2002.

[4] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 1996.

[5] B. Schneier, *Applied Cryptography: protocol, algorithms, and source code in C*, John Wiley & Sons, 2nd Ed., 1996.

[6] J. Li, C. Papodopoulos, J. M. Xu, and M. Moskovits, ", *Applied Physics Letter*, vol. 75, pp. 367–, 1999.

[7] C. Papadopoulos, A. Rakitin, J. Li, A. S. Vedeneev, and J. M. Xu, ", *Physics Review Letters*, vol. 85, pp. 3476–3479, 2000.

[8] R. Martel *et al.*, ", *Phys. Review Letter*, vol. 87, 2002.

[9] C. Zhou, "Electronic transport in carbon nanotubes and molecular scale devices", *NNUN Abstracts 2002/Materials, Physics, Processes and Characterization*, p. 66, 2002.

[10] T. H. LaBean and S. A. Kauffman, "Design of synthetic gene libraries encoding random sequence proteins with desired ensemble characteristics", *Protein Science*, pp. 1249–1254, 1993.

[11] A. P. Blanchard, R. J. Kaiser, and L. E. Hood, "High-density oligonucleotide arrays", *Biosensor, Bioelec.*, vol. 11, pp. 687–690, 1996.

[12] A. Suyama, "DNA chips – integrated chemical circuits for DNA diagnosis and DNA computers", 1998.

[13] D. Kahn, *The codebreakers: the story of secret writing*, New York: Macmillan Pub. Comp, 1967.